Artificial Intelligence Threat Reporting and Incident Response System

D2.3 – Ethics and data protection requirements specification

Project Title:	Artificial Intelligence Threat Reporting and Incident Response System
Project Acronym:	IRIS
Deliverable Identifier:	D2.3
Deliverable Due Date:	30/04/2022
Deliverable Submission Date:	29/04/2022
Deliverable Version:	V1.0
Main author(s) and Organisation:	CyberEthics Lab.
Work Package:	WP2 – System Co-Design
Task:	Task 2.3 – Ethics, legal and societal requirements specification
Dissemination Level:	PU: Public

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no 101021727. Content reflects only the authors' view and European Commission is not responsible for any use that may be made of the information it

Quality Control

	Name	Organisation	Date
Peer Review 1	Susana Zarzosa	ATOS	22/04/2022
Ethics Advisory Board	Teresa Numerico	Ethics Advisory Board	23/04/2022
Peer Review 2	Filippo Rebecchi	THALES	28/04/2022
Submitted by (Project Coordinator)	Nelson Escravana Gonçalo Cadete	INOV	29/04/2022

Contributors

Organisation
CEL
INOV

Document History

Version	Date	Modification	Partner
0.1	01/01/2022	Table of Contents	CEL
0.2	01/03/2022	Internal draft version	CEL
0.3	31/03/2022	Draft version for internal peer review	CEL
1.0	29/04/2022	Final version	CEL, INOV

Legal Disclaimer

IRIS is an EU project funded by the Horizon 2020 research and innovation programme under grant agreement No 101021727. The information and views set out in this deliverable are those of the author(s) and do not necessarily reflect the official opinion of the European Union. The information in this document is provided "as is", and no guarantee or warranty is given that the information is fit for any specific purpose. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein. The IRIS Consortium members shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials subject to any liability which is mandatory due to applicable law.

Contents

List of Abbreviations and Acronyms

Abbreviation / Acronym	Meaning
AI	Artificial Intelligence
CEL	CYBERETHICSLAB Srls
CLS	CYBERLENS BV
CERT	Computer Emergency Response Team
CSIRT	Computer Security Incident Response Team
CERTH	ETHNIKO KENTRO EREVNAS KAI TECHNOLOGIKIS ANAPTYXIS
ECSO	EUROPEAN CYBER SECURITY ORGANISATION
EU	European Union
DPA	Data Protection Authority
DPIA	Data Protection Impact Assessment
FVH	FORUM VIRIUM HELSINKI OY
GDPR	General Data Protection Regulation
ICCS	INSTITUTE OF COMMUNICATION AND COMPUTER SYSTEMS
ICT	Information and Communication Technologies
INOV	INOV INSTITUTO DE ENGENHARIA DE SISTEMAS E COMPUTADORES, INOVACAO
KEMEA	KENTRO MELETON ASFALEIAS

Executive Summary

The present deliverable aims to specify the requirements related to ethics, data protection and secure sharing of data to ensure that the IRIS approach and technology enablers are compliant with the relevant legislative frameworks (i.e., GDPR and NIS Directive) and guidelines (i.e., EU Guidelines for Trustworthy AI). Those requirements have the aim of provide the ethics by design support to the IRIS technology, through the identification of legal and ethics constraints, potential risks raised by those constraints and the requirements to apply the proper safeguards.

The deliverable is implemented in parallel with the definition of the IRIS software system architecture, reflecting the current stage of the project. Eventual requirements modifications or additions will be considered in the next versions of the project architecture deliverables.

1 INTRODUCTION

1.1 Relation to project work

IRIS integrates ground-breaking technologies for automated threat analytics, detection, response and recovery under a novel, domain-specific incident response platform addressed to CERTs/CSIRTs. This platform constitutes IRIS's answer to the challenges which come with (i) proactively managing emergent cybersecurity threats across IoT and AI-driven ICT systems; (ii) sharing and orchestrating threat intelligence effectively and where required at machine-speed among networks of CERTs/CSIRTs; and (iii) training among versatile and diverse cybersecurity teams.

The current deliverable is part of the requirements specification needed to design the IRIS overall system architecture. Therefore, the present document shall not fail to consider its complementarity with T2.2 and related deliverable D2.2 as input for T2.5 and associated deliverables (i.e., D2.5 and D2.6).

Table 1: Relation to other project documents

#ID	Deliverable name	Deliverable description	Due date
D2.2	User and technical requirements	Report on the user & technical requirements that the IRIS platform will have to satisfy	M6
D2.5	IRIS platform and reference architecture – initial version	It will document the initial version of the IRIS reference architecture as well as the technical specifications of the individual IRIS components	M9
D2.6	IRIS platform and reference architecture – final version	It will document the final version of the IRIS reference architecture as well as the technical specifications of the individual IRIS components	M18

1.2 Structure of the document

The document is divided into the following sections:

Table 2: Document structure

No.	Section title	Brief summary
1	Introduction	Provides a brief explanation on the objectives of the IRIS Project, the present deliverable and on the structure of the present document
2	Methodology	Illustrates the methodological approach for carrying out the requirements elicitation process
3	Constraints, risks and requirements on privacy and data protection	Identifies legal privacy and data protection constraints, derived risks and requirements for the implementation of the IRIS technology

No.	Section title	Brief summary
4	Constraints, risks and requirements on ethics and social aspects	Identifies ethics and social aspects constraints, derived risks and requirements for the implementation of the IRIS technology
5	Constraints, risks and requirements on security	Identifies legal security constraints, derived risks and requirements for the implementation of the IRIS technology
6	Conclusions	Summarises the findings of the analysis illustrated in the previous sections

2 METHODOLOGY

The ethics requirements definition methodology applied to this document considers four different steps:

1. Considering that the IRIS software architecture will make use of technologies such as AI cyber threats intelligence ones as well as it will develop a platform for communicating and sharing incident response and recovery, the current document includes a preliminary analysis of the relevant ethics and legal framework principles and constraints that might impact the IRIS Project, including:
   - the protection of personal data processed by the IRIS platform, as provided within the *EU General Data Protection Regulation no. 2016/679*[1] (GDPR);
   - the *Ethics guidelines for Trustworthy AI*[2] the IRIS AI system components should comply with in order to be deemed trustworthy;
   - the adoption of security mechanisms based on the *Directive on security of network and information systems*[3] (NIS Directive) and its revision proposal (NIS2 Directive)[4].

   It is worth to notice that the NIS Directive and the GDPR are linked, meaning that the first one covers some of the data protection requirements of the second one. While the GDPR addresses EU citizen data privacy and how organisations process personal data, the NIS 2 is focused on cyber-risk mitigation using a risk management approach.

2. From this preliminary analysis, three categories of requirements are identified (i.e., privacy and data protection, ethics and social aspects, and security). For each one of them, relevant constraints guiding the IRIS technology implementation are derived for each one of the, including a description and some first recommendations for the Consortium.

3. For each one of the requirements categories, the identified constraints lead to the definition of risks and requirements to be implemented in order to mitigate those risks.

[1] https://eur-lex.europa.eu/eli/reg/2016/679/oj

[2] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[3]https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.194.01.0001.01.ENG&toc=OJ:L:2016:194:TOC

[4] On 3 December 2021, the European Council agreed on its position on the proposal for a Directive on measures for high common level of cybersecurity across the Union (the "NIS2 Directive") https://data.consilium.europa.eu/doc/document/ST-14337-2021-INIT/en/pdf?utm_source=dsms-auto&utm_medium=email&utm_campaign=Strengthening%20EU-wide%20cybersecurity%20and%20resilience%20–%20Council%20agrees%20its%20position

4. Requirements defined in the previous three categories are finally grouped and merged - when they are common for all of them (e.g., requirements related to security mechanisms) – to provide a list of ethics requirements for the IRIS technology implementation.

3 CONSTRAINTS, RISKS AND REQUIREMENTS ON PRIVACY AND DATA PROTECTION

The analysis carried out from the privacy and data protection standpoint is based on the EU current legal framework, namely the General Data Protection Regulation (GDPR). In order to make the reading more effective, the following table summarises the most relevant terms related to the personal data concept adopted by the GDPR.

Term	Definition
Personal Data	Any information relating to an identified or identifiable natural person. In particular an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person
Sensitive Data	Sensitive Data are Personal Data that reveal racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation
Data Processing	Any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction
Profiling	Any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements
Technical and Organizational measure to protect personal data	Any measure designed and implemented to ensure the protection and security of the personal data collected by a controller and/or processor
Pseudonymisation	Means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person

Term	Definition
Controller	Means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data
Processor	Means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller
Data transfer	Means any activities that entail giving access, sharing, transferring or otherwise making available personal data collected/processed by a controller or a processer to another controller or processor

In general terms, it is possible to say that the main privacy and data protection concerns regard a general dis-respect of the following principles and constraints raised by current EU relevant data protection framework, based on the GDPR.

The following table illustrates what are those IRIS relevant constraints stated by the GDPR, including a description and some first recommendations for the Consortium.

Table 3: GDPR constraints

# ID	GDPR constraint	Description
PC1	Transparency	The **purposes of the data processing** should appear clear and intelligible for the data subject. This can be ensured providing all the appropriate and necessary information to data subjects to exercise their rights, to data controllers to evaluate their processors, and to Data Protection Authorities to monitor according to responsibilities. The technology solutions, and their relative **data models**, thus should ensure that a data subject might get easily access, at any time also after the start of the data processing operations, to that information. For the sake of clarity, it should be noted that all that information should be made available to the data subjects in a clear and intelligible way
PC2	Lawful data collection	The data processing should originate from those personal data that have been collected with a lawful ground. Particular attention should be paid when implementing those components that will help to collect and get the **data subject's consent**. In this respect, the relevant Partner should ensure the possibility to map the data flow. Particular attention should be given in case of secondary processing (even if, at the time of submission, this kind of operations are not foreseen)
PC3	Personal data collected are (i) adequate, (ii) proportionate and (iii) relevant to the objectives of the system	The implementation of the principle of **purpose limitation and data minimization**, representing two of the core principles set forth in GDPR, requires that the amount of data collected should be proportionate to the purposes to be achieved, and at the same time, the purpose itself should be legitimate. In this respect, data should be gathered if and only if it is strictly necessary for achieving the specified purpose and that data is "**need to know**"

# ID	GDPR constraint	Description
PC4	The personal data collected are accurate	Besides the amount and the relevancy of the data collected, the technology solutions should ensure that the data to be processed are **accurate**, i.e., data are correct and up-to-date in all details
PC5	Storage Limitation	The development team of the technology solutions should define and implement an infrastructure pursuant to which it is possible to foresee for how long the personal data will be stored (ideally **the shorter the better**), and that in any case shall be compliant with the applicable legislation. Data subjects must be informed about it. Moreover, provided that those data are no longer necessary to fulfil the said scope, and any other restrictions can be found applicable, such data should be immediately erased and/or anonymised pursuant to the best standards and practices
PC6	Procedures for granting individual rights	The components of the technology solutions should be designed taking also into consideration how, in concrete, the relevant data subject might exercise his/her rights in connection with the data processing. In this respect, the relevant Partner should be aware of all the rights that GDPR grants to data subjects, and for each of them tailor **a specific solution** (e.g., data subjects have the right to rectify their data and to request their erasure)
PC7	Accountability principle and technical implementation	The implementation of the accountability principle entails that the technology solutions should allow a clear **identification of the responsibilities** related to the data processing. In particular, examples of accountability measures are related to **tracking** of personal data access and of communications with external systems. In addition, the abovementioned principle implies the set-up of internal audits and handle complaints procedures. Additionally, it should be noticed that at a national level, accountability is supported by independent DPA for monitoring and checking as supervisory bodies

# ID	GDPR constraint	Description
PC8	Implementation of security measures	Information security addresses integrity, confidentiality and availability concerns. PETs (*Privacy Enhancing Technologies - PETs*) represent an important tool (among others) to protect privacy and data protection, in terms of implementing technology solutions able to restrict access to personal data only to authorized people (e.g., **permissions**), and to ensure that the data is trustworthy and accurate (e.g., based on **provenance information**). The relevant Partner should also: (i) regularly conduct privacy risk assessment and audit processes; (ii) regularly run reviews of the security measures implemented; and (iii) design an ad hoc procedure to be followed in case of data breach. Moreover, when it comes to security, besides the principles of confidentiality, integrity and availability, the relevant IRIS Partners should also take into consideration the concepts provided within the ENISA's *Report on Privacy and Data Protection by Design – from policy to engineering*[5] issued in December 2014

In light of this and having in mind the potential data flow within the IRIS architecture and its components, a detailed list of potential risks can be derived. Those risks determine a number of requirements that shall be considered in the IRIS platform architecture development, as defined in the following table:

Table 4: Privacy and Data Protection Risks and Requirements

Req #ID	GDPR constraint	Potential Risk	IRIS Requirement
PR1	PC1 - Transparency	- Data Subject is not informed of (i) which data are collected; (ii) which is the source of the collection; (iii) who are the actors involved in the collection and subsequent processing; and (iv) the purposes of the data processing - Data processing is done for different purposes from the ones agreed with the data subject	- Data exchange shall be carried out if and only if purposes of the data processing is clearly specified in the "contract" among data subject and data controller (i.e., source and destination) - Between the data controller and data processor there shall be a further "contract" to share responsibilities - Purposes of data processing shall be revised at any time, considering changes in data models and purposes of data processing as well

[5] https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design

Req #ID	GDPR constraint	Potential Risk	IRIS Requirement
PR2	PC2 - Lawful data collection	- Data Subject is not aware of data collected and shared - The collection of data is made on a wrong legal basis or in absence of a legal basis	- Data Subject shall be always informed and shall provide consent to data collection and exchange - Data Subject shall always be able to access data to ensure lawfulness and evaluate potential update/rectification - To guarantee the right to be forgotten, data shall be stored in non-DLT storage
PR3	PC3 - Personal data collected are (i) adequate, (ii) proportionate and (iii) relevant to the objectives of the system	Collection of unneeded (personal) data, i.e., data not relevant to the objectives of the system and for the agreed purposes of data processing	- When defining the data model of the component, each single data property shall be strongly justified, by applying the "need-to-know" principle - Data aggregation, anonymization and pseudonymisation techniques shall be adopted for the purpose of component testing, demonstration and operation
PR4	PC4 - The personal data collected are accurate	Lack of information among involved parties is the primary potential cause for inaccurate data in a system	- Data Subjects and Data Controllers shall be continuously informed about the status of the ongoing data sharing activities, as well as of their requests for changes (i.e., fundamental information for ensuring accuracy of exchanged information) - The appropriate interfaces shall be defined and assessed with the continuous engagement of Data Subjects and Data Controllers
PR5	PC5 - Storage limitation	Data persistency has to be guaranteed for the minimum required timeframe, according to contracts among parties and the applicable regulatory framework	According to the purposes of the system, each single component of the IRIS architecture shall contribute to the definition of the minimum storage timeframe. This relevant parameter shall be based on components data model

Req #ID	GDPR constraint	Potential Risk	IRIS Requirement
PR6	PC6 - Procedures for granting individual rights	Lack of information of the data subject rights at design phase impacts on enabling/disabling the exercise of individual rights themselves	Updates in the data model of the components shall be handled to identify potential personal/sensitive data and consequently to plan how components enable/disable the exercise of individual rights (including rectification and/or erasure)
PR7	PC7 - Accountability principle and technical implementation	Accountability of the system is impacted by the lack of provenance information regarding activities of the components (i.e., logs), access to the system, integrity of data collected, integrity of data exchanged	- Adequately trace the data exchange, and integrity of data exchange with appropriate tools and techniques (e.g., log, provenance information, hashing algorithms) - DLT technology, that is going to be considered for the IRIS technology, represents a key contributor for ensuring the traceability and data integrity
PR8	PC8 - Security measures	Risks are highlighted in the section on security (see 5)	The actions required to mitigate these risks are highlighted in the section on security requirements

4 CONSTRAINTS, RISKS AND REQUIREMENTS ON ETHICS AND SOCIAL ASPECTS FOR AI

This section aims to explain the Ethics guidelines for trustworthy Artificial Intelligence (AI) – developed by the High-Level Expert Group on AI, 2019, in relation to the architecture and the solutions based on AI mechanisms proposed by the IRIS project.

The purpose of the guidelines is to outline ethical requirements for the development of trustworthy AI systems. Trustworthiness is a concept articulated through the categories of a) legality, b) ethicality, and c) robustness.

The three dimensions are closely interdependent and it is necessary to develop all of them in a harmonious way in order to talk about trustworthiness of an AI system.

To offer guidance on the implementation and realisation of Trustworthy AI in IRIS, a list of seven requirements should be met (Figure 1). These requirements are applicable to different stakeholders partaking in AI systems' life cycle: developers, deployers and end-users, as well as the broader society. "*By developers, we refer to those who research, design and/or develop AI systems. By deployers, we refer to public or private organisations that use AI systems within their business processes and offer products and services to others. End-users are those engaging with the AI system, directly or indirectly. Finally, the broader society encompasses all others that are directly or indirectly affected by AI systems*" (Guidelines 2019, p. 14).

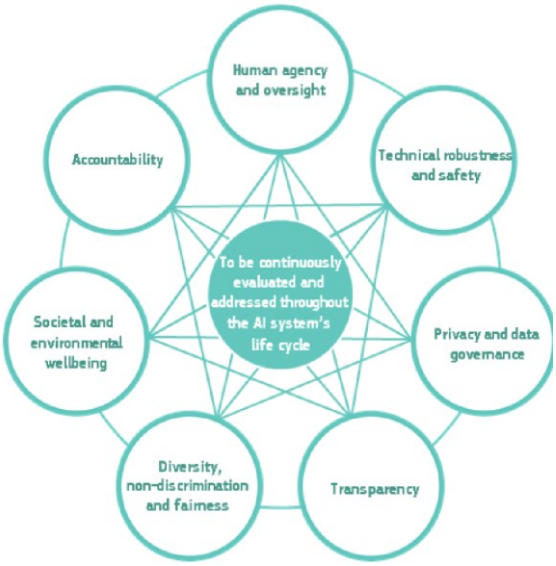

Figure 1: Key requirements (Source: EU Guidelines 2019).

The list of requirements includes systemic, individual and societal aspects:

1. **Human agency and oversight** – *Including fundamental rights, human agency and human oversight.*
2. **Technical robustness and safety** – *Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility.*
3. **Privacy and data governance** – *Including respect for privacy, quality and integrity of data, and access to data.*
4. **Transparency** – *Including traceability, explainability and communication.*
5. **Diversity, non-discrimination and fairness** – *Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation.*
6. **Societal and environmental wellbeing** – *Including sustainability and environmental friendliness, social impact, society and democracy.*
7. **Accountability** – *Including auditability, minimisation and reporting of negative impact, trade-offs and redress*.

The following list will be adopted as the guiding framework to elaborate the related constraints and requirements for the IRIS platform.

Table 5: EU Guidelines for trustworthy AI constraints

# ID	Trustworthy AI constraint	Description
EC1	Human agency and oversight	With the aim of empowering human beings, AI systems should support **human autonomy**, thus enabling informed and autonomous decision-making that respects human agency. In addition, AI systems must be catalysts for democratic values, respecting and promoting human rights and always allowing for **human oversight**. Thus, a rights impact assessment should be carried out prior to the development of the system, and external feedback regarding potential rights violations should always be ensured. Furthermore, to ensure human agency, governance mechanisms such as human-in-the-loop (HITL), human-on-the-loop (HOTL), and human-in-command (HIC) should be implemented.
EC2	Technical Robustness and safety	In order to ensure the trustworthiness and harm prevention of an AI system, it must be robust and resilient to attacks. The concept of robustness is articulated by implementing actions that guarantee a) **Resilience to attacks and security**, considering potential abuse of the system by malicious actors, b) **Fall-back plan**: in case of problems, having a fall-back plan can prevent damage to people, things and the environment, c) **Accuracy**: The AI system must have a high degree of accuracy in making judgements, classifying information, making predictions, d) **Reliability and Reproducibility:** the reliability of an AI system is closely related to the reproducibility of its operation. Reproducibility ensures that the system always works in the same way under the same conditions, thus guaranteeing the predictability and reliability of the system.

# ID	Trustworthy AI constraint	Description
EC3	Privacy and data governance	Since privacy is one of the elements most impacted by AI systems, it is necessary to have tools that ensure proper privacy protection, for which we refer to the requirements outlined by the **GDPR**.
EC4	Transparency	Transparency, a concept linked to explainability, must be guaranteed for various aspects of AI systems: the data and processes that constitute the system must be **documented** and **traceable**, every process (both technical and decision-making) must be **comprehensible** to human beings (whether technicians, researchers or regulators). Finally, humans must always be made **aware** that they are interacting with an AI system and not with other humans. Preventing this kind of deception is extremely important to ensure trustworthy communication.
EC5	Diversity, non-discrimination and fairness	A trustworthy AI system is one that guarantees **equality** and **non-discrimination** throughout the life cycle of the system. For this to be possible, it is necessary to avoid **bias;** to guarantee a **universal design** that ensures accessibility to all users, taking into account the diversity and differences of every human being; and finally, it is essential to actively **involve the stakeholders** who will be directly or indirectly affected by the AI process throughout the system's life cycle, thus also guaranteeing the possibility of feedback even after the system has been developed.
EC6	Societal and environmental well-being	Loyalty, trust and harm prevention should not only be thought of in relation to human beings, but to society as a whole, thus including **all sentient beings** and the environment. This is possible if AI systems are developed by carefully assessing the **environmental and energy impact**, monitoring and considering the **social impacts** on the psycho-physical well-being of all community members and taking into account the impact on the **democratic and participatory processes** of societies.
EC7	Accountability	It is necessary that AI systems, throughout their lifecycle, have mechanisms whereby the **consequences** of actions taken can be accounted for. This is possible if strategies are put in place such as **auditability**: the possibility that auditors can always evaluate the algorithms and processes of the system**; a negative impact reporting**: it is crucial to have impact assessments both before and during the development and use of the AI system in order to minimise risks and negative impacts. It is necessary to document any **trade-offs** that are made if conflicts between risks arise, the decision maker must be accountable for how a trade-off is adopted. Finally, for an AI system to be trustworthy, when a negative and unfair impact occurs, it must be possible to seek appropriate **redress** for the harm suffered.

Having this ethical framework in mind, it is possible to identify what the major risks of not applying the above requirements might be, and it is also possible to identify risk mitigation measures specifically designed for the IRIS platform structure.

Table 6: Trustworthy AI Risks and Requirements

Req #ID	Trustworthy AI constraint	Potential Risk	IRIS Requirement
ER1	EC1 - Human agency and oversight	- The subject is unable to make autonomous and informed choices - Subject's dignity as an agency person is violated	Human in the loop and Human in command mechanisms shall be implemented
ER2	EC2 - Technical Robustness and safety	- The system could be used by malicious actors - In case of damage, if there is no fall-back plan, the damage may extend to things, people, environment - The system may not provide correct and accurate indications and information - If the system does not have a high rate of reproducibility, it may be unpredictable	- Non-repudiation mechanisms shall be implemented - An accurate test plan to be reproduced over time to ensure the efficiency and proper functioning of the system shall be prepared, so that the degree of accuracy and reproducibility can be checked and verified - System stakeholders shall be adequately informed e.g. throw adequate informative material
ER3	EC3- Privacy and data governance	- Risks are highlighted in the section on data protection and governance (see 0)	The actions required to mitigate these risks are highlighted in the section on GDPR requirements (see 0)
ER4	EC4- Transparency	-The system is difficult to explain and understand	As the information processed by the IRIS platform is strictly confidential and relevant to security issues, processes and system behaviour (both technical and decision making) shall be carefully documented and tracked to ensure transparency

Req #ID	Trustworthy AI constraint	Potential Risk	IRIS Requirement
ER5	EC5- Diversity, non-discrimination and fairness	- The presence of discriminatory bias leads to actions that may marginalize and discriminate against certain groups or categories of people - Non-universal design may exclude certain categories of people (e.g. people with disabilities) - If stakeholders are not involved, the system may be developed in an undemocratic way	- Decision-making processes shall not be made based on discriminatory bias. A group of external experts shall be consulted to make assessments and analyses of possible discriminatory biases - The platform interface and functionalities shall be universally accessible to all human beings, respecting their diversity - Co-design involving all relevant stakeholders' categories shall be ensured
ER6	ER6- Societal and environmental well-being	- The system might harm not only people, but also other sentient beings, the environment and the society as a whole - If adequate measures are not taken, the impact of the AI system on the mental and physical well-being of people and the community may not be properly assessed	The system shall be sustainable from an environmental and energetic point of view, , being compliant with the Do Not Significant Harm (DNSH)[6] principle
ER7	EC7- Accountability	- Without appropriate auditability and redress measures, the system might be considered untrustworthy - It might be difficult to trace processes	- A lead manager who is responsible for the AI system who can account for the consequences of actions taken shall be identified and communicated to the stakeholders - A tracking mechanism shall be implemented to log accesses and actions carried out by using the system

[6] https://ec.europa.eu/info/sites/default/files/2021_02_18_epc_do_not_significant_harm_-technical_guidance_by_the_commission.pdf

5 CONSTRAINTS, RISKS AND REQUIREMENTS ON SECURE DATA SHARING

The analysis of the legislative requirements imposed at EU level is expressed in relation to the need of secure sharing of incidents response in case of cyber threats and recovery information, implemented by the IRIS technical solution.

Therefore, the most appropriate legal framework from which it is possible to infer the security requirements that the Project should implement through its components within the IRIS architecture is referred to the NIS Directive and its revision proposal (NIS2 Directive).

The NIS Directive has three main objectives:

1. Improving national cybersecurity capabilities;
2. Building and fostering cooperation (on cybersecurity) at EU level, requiring Member States to elaborate a National Cybersecurity strategy, to establish Computer Security Incident Response Teams (CSIRTs) and to appoint NIS national competent authorities;
3. Promoting a culture of risk management and incidents reporting among key economic actors, operators providing essential services for the maintaining of economic and societal activities, and digital service providers.

The NIS Directive, entered into force in August 2016 and transposed into Member States national laws by 9 May 2018, is the first horizontal piece of legislation aimed at protecting the security of network and information systems.

However, some issues have raised from its application, mainly the difficulty to implement it, resulting in fragmentation at different levels across the internal market, as well as issues related to the growing of threats due to digitalization and the surge of cyberattacks. Thus, the Commission launched on 7 July 2020 – closed on 2 October 2020 - a public consultation on the revision of the NIS Directive that aims to collect views on its implementation and on the impact of potential future changes.

On 16 December 2020, the European Commission and the High Representative of the Union for Foreign Affairs and Security Policy defined a new EU Cybersecurity Strategy, presenting two new proposals: a Directive on measures for high common level of cybersecurity across the Union to repeal the existing NIS Directive (with the so called NIS2 Directive), and a new Directive on the resilience of critical entities.

The new NIS2 Directive expands the NIS Directive scope, aiming to strengthen the security requirements imposed, addressing security of supply chains, streamlining reporting obligations, introducing more stringent supervisory measures and stricter enforcement requirements including harmonised sanctions regimes across Member States. It also includes proposals for information sharing and cooperation on cyber crisis

management at national and EU level. The proposed expansion of the scope covered by the NIS2 would effectively oblige more entities and sectors to take measures, increasing the level of cybersecurity in the EU longer term.[7]

On 13 April 2021 the Commission presented its proposal, and on 26 May 2021 the draft report was delivered.

The final report was adopted by Parliament in its plenary of 22 November 2021 together with the decision to enter into interinstitutional negotiations.

The Council adopted its negotiating position on 3 December 2021, introducing a number of significant changes to the Commission's proposal and trying to align it with other related proposed legislation, such as the Directive on the resilience of critical entities (CER Directive) and the proposed Regulation on digital operational resilience for the financial sector (DORA).

The main changes are as follows:

- expanding the category of essential entities: in the NIS Directive, covered entities were defined as "operators of essential services" (OESes) and "digital services providers" (DSPs). In the new directive, the current distinction between digital service providers and operators of essential services is eliminated, so that the entities are categorized in 'essential' or 'important', depending on the organisation's criticality in terms of the economy and society. Eight key industry sectors are covered by the NIS2, while it excludes entities operating in defence or national security, public security, law enforcement and the judiciary, as well as parliaments and central banks;
- simplifying the incident reporting obligations to avoid over-reporting, excluding the mandatory reporting for significant cyber threats to the competent authorities or the Computer Security Incident Response Teams (CSIRT);
- clarifying jurisdiction for entities based on their type;
- establishing a European cyber crises liaison organisation network (EU-CyCLONE) to support the coordinated management of large-scale cybersecurity incidents and crises at EU level;
- adding a risk assessment approach in line with other legal frameworks such as for instance the GDPR. Such a risk assessment as well as the incident response should lead to the implementation of security measures outlined in the directive. The level of requirement for cybersecurity risk management and reporting obligations depends on the 'important' or 'essential' entity classification assigned to an organization. Moreover, more stringent risk management is required on the

[7] https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-review-of-the-nis-directive

entire supply chain. Finally, the concept of accountability is reinforced applying to the whole organization and not only to the IT function;

- strengthening the security requirements for the companies subject to the rules by providing a minimum list of basic compulsory security elements and introducing more precise incident response reporting requirements. Included security measures are:
  - Risk analysis and information system security policies,
  - Business continuity and crisis management,
  - Vulnerability handling and disclosure,
  - Cyber security testing and auditing,
  - Effective use of encryption,
  - Multi-factor authentication,
  - Secured voice, video, and text communications,
  - Secured emergency communications systems.

  In particular, new security constraints related to the human resources security (access control policies have been newly introduced by the Council's proposal accordingly with the CER Directive[8] on the resilience of critical entities) are included;

- reinforcing the proportionality principle with regard to the technical and organisational measures, that shall take into account the degree of the entity's exposure to risks, its size, the likelihood of occurrence of incidents and their severity;

- boosting a higher degree of harmonisation at Union level, through an obligation of the Commission to adopt an implementing act that should facilitate the implementation of cybersecurity measures and include certain entities (e.g., cloud computing service providers, data centre service providers, content delivery network, and trust service providers);

- extending the period for Member States to transpose NIS2 into national law to two years, instead of 18 months.

Trilogue interinstitutional negotiations started on 13 January 2022.

5.1 The NIS Directive currently in force

Since the trilogue interinstitutional negotiations will take some time and considering that Member States will still have 24 months to transpose the directive into their national laws once the final text is agreed on, it is expected that the new directive will not repeal the current one until the 2024.

[8] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A829%3AFIN

In the meantime, the current analysis will take into account the still in force NIS Directive provisions, keeping open the possibility of updating it when the NIS2 will enter into force.

The NIS Directive has established a Cooperation Group[9] to ensure cooperation and information exchange among the Member States. The Group aims to achieve a high standard of security for network and information systems in the European Union by supporting and facilitating strategic cooperation and exchange of information among EU Member States and by providing several non-binding guidelines to the EU Members States to allow effective and coherent implementation of the NIS Directive.

On the operational side, the NIS Cooperation Group is supported by the work of the network of Computer Security Incident Response Teams (CSIRTs), dedicated to sharing information about risks and ongoing threats, and cooperating on specific cybersecurity incidents. The CSIRTs network was established under Article 12 of the NIS Directive, which also defines its role. The NIS Cooperation Group provides strategic guidance for the activities of the CSIRTs network.

The NIS Cooperation Group identifies the following 3 macro – areas (each of them sub-categorized) in which specific security policies should be implemented in the following table.

Table 7: NIS Cooperation Group macro-area

GDPR constraint	Description
Governance and Ecosystem	Information System Security Governance & Risk Management
	Ecosystem management
Protection	IT Security Architecture
	IT Security Administration
	Identity and Access management
	IT Security maintenance
	Physical and environmental security
Defence	Detection
	Computer security incident management

In particular the NIS Directive defines an incident as "*any event having an actual adverse effect on the security of network and information systems*". In order to determine the significance of the impact of an incident, operators of essential services and digital service providers must take into account the following parameters:

1. the number of users affected by the disruption of the essential service;
2. the duration of the incident; and
3. the geographical spread with regard to the area affected by the incident.

In terms of compliance with NIS Directive set of obligations, the same can be distinguished in obligations to (i) notify the national legislator/regulators concerning

[9] https://digital-strategy.ec.europa.eu/en/policies/nis-cooperation-group

incidents that met a certain threshold, and (ii) voluntary disclose information/incidents[10], which, according to the NIS Cooperation Group Guidelines on notification of Operators of Essential Services (OES) incidents – Formats and procedures" publication 05/2018, can allow authorities to get a better situational awareness as well as to identify potential new threats and consequently informs also other OES.

As per the notification obligations, the NIS Cooperation Group published a useful guideline in May 2018, aimed at providing non-binding technical guidance "*to national competent authorities and CSIRTs with regard to formats and procedures for the notification of incidents by OES, to facilitate alignment in the implementation of the NIS Directive across the EU*".[11] Indeed also in this case the adoption of uniform guidelines could represent a vital asset to tackle cross-border incidents, improve collaboration and the aggregation of the data and their analysis, as well as improve the entire efficiency of the system.

In particular, in terms of notification procedures, the NIS Cooperation Group states that the timing of the notification will have to take place without unjustified delay. It also provides the following:

- alert notifications to be addressed to the competent national authority or to the competent Computer Security Incident Response Team (CSIRT) in order to:
  - "*Offer support to the affected organization, for example, the CSIRT could give technical support.*
  - *Assess the potential impact for essential services, citizens, the society, the economy, etc.*
  - *Inform, in exceptional circumstances, and when this is in the public interest, other organizations, so they can take action.*
  - *Prevent spreading or reduce the impact by warning and sharing information with relevant organizations, for example with other OESs, CSIRTs, etc.*
  - *Inform authorities abroad when there is significant impact across the EU*"[12].
- Follow up notifications to update on the status of the alert notification.

In addition, the document highlights how much is important the timing of the notification itself, proposing also different methods to transmit the same, as well as indicating that the same notifications shall be also protected.

As far as the information sharing, the NIS Cooperation Group published some guidelines in January 2019,[13] stating the voluntary nature of notification, based on a three steps framework as follows:

[10] Michels, Johan David and Walden, Ian, How Safe is Safe Enough? Improving Cybersecurity in Europe's Critical Infrastructure Under the NIS Directive (December 7, 2018). Queen Mary School of Law Legal Studies Research Paper No. 291/2018. Available at SSRN: https://ssrn.com/abstract=3297470

[11] NIS Cooperation Group "Guidelines on notification of Operators of Essential Services incidents – Formats and procedures" publication 05/2018

[12] Ibidem, page 11

1. *First step – who should start the dialogue and why: Information exchange[14] on cross-border dependencies shall be conducted by SPOCs [Single Point of Contact] of Member States as responsible authorities for coordinating issues related to security of network and information systems[15] as the SPOC designated under the NIS Directive is considered as a key national entity to undertake the[16] information exchange and liaison function on behalf of each Member State.*

2. *Second step – content of the dialogue: AMS [Affected Member State] will need to provide the following information to OMS [Originating Member State]:*

- *Description of the service or network and information system in OMS, upon which an essential service in AMS is dependent upon.*
- *Description of the service provider (Operator of essential service) in AMS*
- *Questions related to network and information security of the service in OMS that the essential service of AMS is dependent upon and that AMS needs more information about in order to support its national risk management process. These questions may notably include information about security measures or requirements of network and information security that are in place for the given service or network and information system. For example about possible measures or requirements related to service continuity like Maximum Tolerable Downtime (MTD) or Recovery Time Objective (RTO).*

3. *Third step - Based on the received information the AMS can:*

- *Establish further discussion with the OMS SPOC on possibilities for mitigating identified dependencies*
- *Establish additional risk mitigation measures within AMS, taking into account the results of the dialogue.*

On the basis of the abovementioned legal framework, the following security constrains are illustrated in the following table:

[13] Guidelines for the Member States on voluntary information exchange on cross-border dependencies, CG Publication 01/2019
https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=65182

[14] It is important to take into account that Member States can have special requirements around information sharing and they therefore might require assurances for channels of information sharing, which may also be be delayed due to internal clearance processes.

[15] NIS Directive, rec.31

[16] According to article 8(4) of the Directive, the single point of contact shall exercise a liaison function to ensure cross-border cooperation of Member State authorities and with the relevant authorities in other Member States and with the Cooperation Group and the CSIRTs network. However, this does not preclude a Member State from choosing national authorities other than SPOCs and national competent authorities under the Directive to undertake this task.

Table 8: NIS Directive constraints

# ID	NIS constraint	Description
SC1	Implementation of security measures	The IT platform has to implement adequate and appropriate security measures able to protect the data to be ingested in the platform as well as its functionalities. In this respect, such measures shall include either hardware measures as well as software ones, and in any case shall be designed applying a risk-based approach, which has to consider all the components and their interactions
SC2	Notification system	The platform has to be able to (i) detect and to send a prompt warning notification/message in case of actual attacks or even potential to the most appropriate authority; (ii) send a notification message complete with all the necessary information to detect the threats and determine the countermeasures; and (iii) the same notification system has also to be designed and construed applying adequate and proportionate security measures
SC3	Information sharing	The information sharing system should provide at least the following info: • description of the service or network and information system in OMS, upon which an essential service in AMS is dependent upon. • description of the service provider (Operator of essential service) in AMS • questions related to network and information security of the service in OMS that the essential service of AMS is dependent upon and that AMS needs more information about in order to support its national risk management process. These questions may notably include information about security measures or requirements of network and information security that are in place for the given service or network and information system. It should support the OMS SPOCs to • mitigate identified dependencies • mitigate additional risks.
SC4	Confidentiality	Both personal and non-personal information have to be protected from un-authorized access and/or use
SC5	Availability	The information circulating within the IT system have to be timely and reliably accessible in case of need
SC6	Integrity	The information stored or in any case circulating within the IT platform cannot be modified (nor be tampered or loss), and therefore have to be reliable and trustable
SC7	Accountability	The information (i.e., data) and the operations made on certain data can be tracked and traced back to specific and pre-authorised individuals. Ensuring the respect of the accountability therefore entails the respect of the principle of authenticity

Having in mind the abovementioned table, it is also possible to identify a series of potential threats or potential risks and requirements in the following table.

Table 9: Cyber Security Risks and Requirements

Req #ID	NIS constraint	Potential Risk	IRIS Requirement
SR1	SC1 - Implementation of security measures (in general)	- Appropriate security measures either at organizational and at technical level have not been developed/have been wrongly implemented. In particular, there might be the risk to cover all the identified potential threats but the implementations are not sufficiently flexible to cover also unforeseen events - An alignment among the security measures *strictu sensu* and the security measures implemented to ensure the privacy and data protection rights has not been performed and such dis-homogeneity might create conflicts	- Security test procedures, acceptance thresholds and reports shall be specified in order to evaluate the addressing of all the defined threats, as well as to identify new potential and unforeseen threats. - IRIS components shall be delivered with relative test reports, in order to provide evidence of security level
SR2	SC2 - Notification system	The system has not been designed to provide timely alerts and/or the addressee of the alerts have not been correctly identified, or the alert chain is per se not secured and possible intrusions or interferences might happens jeopardising the alert system itself and the messages contained	- Parties (i.e., data subject and data controller) shall be promptly notified about the status of any event occurred in the system and that can directly or indirectly impact on them - Notification system shall adopt appropriate measures in order to guarantee the authenticity and integrity of alerts themselves

Req #ID	NIS constraint	Potential Risk	IRIS Requirement
SR3	SC3 - Information sharing	The information sharing system doesn't provide the minimum set of information needed and requested by low. The information sharing system doesn't support the OMS SPOCs to mitigate identified dependencies and additional risks	The IRIS system design and implementation shall be based on co-creation methodologies fostering a strict collaboration between all stakeholders, including a risk assessment approach and a feedback loop to ensure flexibility and prompt reaction to changes
SR4	SC4 - Confidentiality	An improper definition and management of authorisations to access and/or use data might entails: (i) several vulnerabilities and impact on the confidentiality of its managed information; (ii) the violation of several GDPR provisions	- Appropriate management of authorisations shall be ensured to access and/or use data - The level of reputation of the entities involved to gather, collect, access and process data shall be continuously monitored. Based on the updated information, authorisation to access and/or use data shall be accordingly revised
SR5	SC5 - Availability	Overload of security operations might potentially impact on timely access to important information, necessary for the proper operating conditions of the smart grid	A reasonable level of security shall be identified with respect to the time constraints. Lightweight hashing algorithms and performing encryption mechanisms shall be considered at the design phase of the communication protocols and mechanisms of the architecture
SR6	SC6 - Integrity	Data might undergo several transformations (e.g., format and protocol) impacting on its authenticity and integrity	Any operation on data (including the authorised permissions) shall be tracked in a secure and trustable register, in order to provide the evidences of integrity and authenticity of data managed by the system

Req #ID	NIS constraint	Potential Risk	IRIS Requirement
SR7	SC7 - Accountability	If any specific data transformation is performed without ensuring the traceability of authorised permissions, or permissions are not assigned to trustable entities, accountability of the system is definitely compromised, as well as the authenticity of its managed information	Adequate technology shall be adopted for ensuring the traceability of permissions, authorisations, reputations, events, and any vital information needed for providing evidence of system accountability, and data authenticity and integrity

6 CONCLUSIONS

The analysis carried out in the previous sections concludes with the definition of the ethics requirements corresponding to the constraints derived from the project relevant EU legal and guidelines framework.

The complete analysis is then illustrated in the following tables, where the requirements have an *ID*, a *name* and a *rationale*, referring to the specific constraints identified in the previous sections, as well as a *description* summarising the actions that the partners should implement into the IRIS technology. Since the requirements are linked to constraints, they are all mandatory for a compliant implementation of the IRIS technology (i.e. "Priority" = "MUST").

ID	Ethics_01	Priority	MUST
Name	*Transparency*		
Description	- Data exchange shall be carried out if and only if purposes of the data processing is clearly specified in the "contract" among data subject and data controller (i.e., source and destination) - Between the data controller and data processor there shall be a further "contract" to share responsibilities - Purposes of data processing shall be revised at any time, considering changes in data models and purposes of data processing as well - As the information processed by the IRIS platform is strictly confidential and relevant to security issues, processes and system behaviour (both technical and decision making) shall be carefully documented and tracked to ensure transparency		
Rationale	Privacy constraint PC1 Ethics and social for AI constraint EC4		

ID	Ethics_02	Priority	MUST
Name	*Lawful data collection*		
Description	- Data Subject shall be always informed and shall provide consent to data collection and exchange - Data Subject shall always be able to access data to ensure lawfulness and evaluate potential update/rectification - To guarantee the right to be forgotten, data shall be stored in non-DLT storage		
Rationale	Privacy constraint PC2		

ID	Ethics_03	Priority	MUST
Name	*Personal data collected are (i) adequate, (ii) proportionate and (iii) relevant to the objectives of the system*		
Description	- When defining the data model of the component, each single data property shall be strongly justified, by applying the "need-to-know" principle - Data aggregation, anonymization and pseudonymisation techniques shall be adopted for the purpose of component testing, demonstration and operation		
Rationale	Privacy constraint PC3		

ID	Ethics_04	Priority	MUST
Name	*The personal data collected are accurate*		
Description	- Data Subjects and Data Controllers shall be continuously informed about the status of the ongoing data sharing activities, as well as of their requests for changes (i.e., fundamental information for ensuring accuracy of exchanged information) - The appropriate interfaces shall be defined and assessed with the continuous engagement of Data Subjects and Data Controllers		
Rationale	Privacy constraint PC4		

ID	Ethics_05	Priority	MUST
Name	*Storage limitation*		
Description	According to the purposes of the system, each single component of the IRIS architecture shall contribute to the definition of the minimum storage timeframe. This relevant parameter shall be based on components data model		
Rationale	Privacy constraint PC5		

ID	Ethics_06	Priority	MUST
Name	*Procedures for granting individual rights*		
Description	Updates in the data model of the components shall be handled to identify potential personal/sensitive data and consequently to plan how components enable/disable the exercise of individual rights (including rectification and/or erasure)		
Rationale	Privacy constraint PC6		

ID	Ethics_07	Priority	MUST
Name	*Accountability principle and technical implementation*		
Description	- Adequately trace the data exchange, and integrity of data exchange with appropriate tools and techniques (e.g., log, provenance information, hashing algorithms) - DLT technology, that is going to be considered for the IRIS technology, represents a key contributor for ensuring the traceability and data integrity - A lead manager who is responsible for the AI system who can account for the consequences of actions taken shall be identified and communicated to the stakeholders - A tracking mechanism shall be implemented to log accesses and actions carried out by using the system - Adequate technology shall be adopted for ensuring the traceability of permissions, authorisations, reputations, events, and any vital information needed for providing evidence of system accountability, and data authenticity and integrity		
Rationale	Privacy constraint PC7 Ethics and social for AI constraint EC7 Security constraint SC7		

ID	Ethics_08	Priority	MUST
Name	*Human agency and oversight*		
Description	Human in the loop and Human in command mechanisms shall be implemented		
Rationale	Ethics and social for AI constraint EC1		

ID	Ethics_09	Priority	MUST
Name	*Technical Robustness and safety*		
Description	- Non-repudiation mechanisms shall be implemented - An accurate test plan to be reproduced over time to ensure the efficiency and proper functioning of the system shall be prepared, so that the degree of accuracy and reproducibility can be checked and verified - System stakeholders shall be adequately informed e.g. throw adequate informative material		
Rationale	Ethics and social for AI constraint EC2		

ID	Ethics_10	Priority	MUST
Name	*Diversity, non-discrimination and fairness*		
Description	- Decision-making processes shall not be made based on discriminatory bias. A group of external experts shall be consulted to make assessments and analyses of possible discriminatory biases - The platform interface and functionalities shall be universally accessible to all human beings, respecting their diversity - Co-design involving all relevant stakeholders' categories shall be ensured		
Rationale	Ethics and social for AI constraint EC5		

ID	Ethics_11	Priority	MUST
Name	*Societal and environmental well-being*		
Description	The system shall be sustainable from an environmental and energetic point of view, being compliant with the Do Not Significant Harm (DNSH) principle		
Rationale	Ethics and social for AI constraint EC6		

ID	Ethics_12	Priority	MUST
Name	*Implementation of security measures*		
Description	- Security test procedures, acceptance thresholds and reports shall be specified in order to evaluate the addressing of all the defined threats, as well as to identify new potential and unforeseen threats. - IRIS components shall be delivered with relative test reports, in order to provide evidence of security level		
Rationale	Privacy constraint PC8 Security constraint SC1		

ID	Ethics_13	Priority	MUST
Name	*Notification system*		
Description	- Parties (i.e., data subject and data controller) shall be promptly notified about the status of any event occurred in the system and that can directly or indirectly impact on them - Notification system shall adopt appropriate measures in order to guarantee the authenticity and integrity of alerts themselves		
Rationale	Security constraint SC2		

ID	Ethics_14	Priority	MUST
Name	*Information sharing*		
Description	The IRIS system design and implementation shall be based on co-creation methodologies fostering a strict collaboration between all stakeholders, including a risk assessment approach and a feedback loop to ensure flexibility and prompt reaction to changes		
Rationale	Security constraint SC3		

ID	Ethics_15	Priority	MUST
Name	*Confidentiality*		
Description	- Appropriate management of authorisations shall be ensured to access and/or use data - The level of reputation of the entities involved to gather, collect, access and process data shall be continuously monitored. Based on the updated information, authorisation to access and/or use data shall be accordingly revised		
Rationale	Security constraint SC4		

ID	Ethics_16	Priority	MUST
Name	*Availability*		
Description	A reasonable level of security shall be identified with respect to the time constraints. Lightweight hashing algorithms and performing encryption mechanisms shall be considered at the design phase of the communication protocols and mechanisms of the architecture		
Rationale	Security constraint SC5		

ID	Ethics_17	Priority	MUST
Name	*Integrity*		
Description	Any operation on data (including the authorised permissions) shall be tracked in a secure and trustable register, in order to provide the evidences of integrity and authenticity of data managed by the system		
Rationale	Security constraint SC6		